Packetisation in Optical Packet Switch Fabrics using adaptive timeout values

Mortensen, Brian Bach

Published in:
High Performance Switching and Routing, 2006

Link to article, DOI:
10.1109/HPSR.2006.1709710

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Packetisation in Optical Packet Switch Fabrics using adaptive timeout values

Brian B. Mortensen
COM·DTU
Technical University of Denmark
DK-2800 Kgs. Lyngby
Email: bbm@com.dtu.dk

***Abstract*—Hybrid electro-optical packet switches utilize optics in the backplane to switch optical packets from inputs to outputs on electronic line cards. The optical packets are traditionally considerably larger than minimum size IP packets. IP packets entering the switch must be formatted (segmented) and encapsulated in the optical packet format. This process is called packetisation or aggregation. This paper investigates a novel technique for aggregating IP packets into optical packets. When the first segment arrives in an optical packet, a timer is started. The optical packet is marked ready for transmission either because the timer reaches a specific timeout value, or because the optical packet is completely filled with segments. Only two distinct values of the timeout value are used. Which of the two timeout values to use, is selected by 3 different control thresholds. The first threshold level applies to the inter arrival rate at the individual VOQs. The remaining thresholds applies to the optical slot level inter arrival rate at the input and output line cards. If any measurements are beyond a given threshold, the higher timeout value is used. The proposed method can be used to make a trade-off between delay and throughput in hybrid electro-optical packet switching. Furthermore, it is investigated how large a speedup is required in order to provide 100% throughput.**

***Index Terms*— Adaptive Timeouts, Bufferless Crossbar, Optical Packet Switching (OPS), Packet Aggregation.**

I. Introduction

USING optics to build large scale packet switches (>1Tb/s) has been proposed in the literature due to potential advantages regarding power consumption and scalability [1] [2]. Using optics in the backplane makes it possible to avoid expensive conversion of signals from electrical to optical domain and vice versa. This architecture is shown in Figure 1. Besides reducing component cost there is a potential for saving power in the interconnection links, which traditionally consumes a large amount of power in electrical packet switches. Optical packet switches with a total capacity of 2.56Tb/s have been presented in the literature [3]. This type of optical packet switch architecture utilizes the broadcast and select scheme where all wavelengths are amplified and broadcasted to a space and wavelength selections unit. The space and wavelength selection unit utilizes SOA (Semiconductor Optical Amplifiers) gates, to select a wavelength from a specific input port. The optical packet switch (OPS) in [3] was operated in a time slotted scheme, thus employing fixed length optical packets.

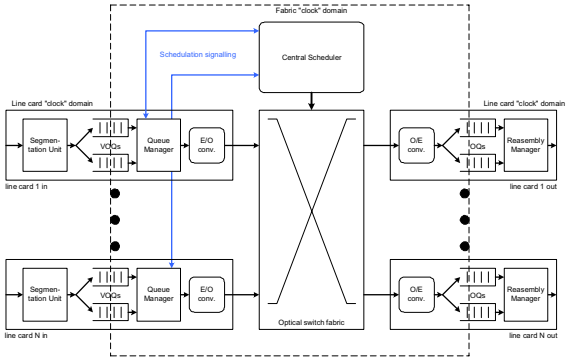

Figure 1 Hybrid Electro-Optical Packet switch. Ingress part of the line cards utilizes Virtual Output Queuing (VOQ). Egress part of line card uses output queuing structure to reassemble segments from different inputs. A possible speedup between core frequency and line card is illustrated by the dashed box.

The time slotted scheme implies a guard band between optical packets to allow for reconfiguration of the optical switch matrix. Furthermore, overhead is needed in the optical packet in the form of a delimiter, which can be used by the receiver to recover the data. This is needed even in the case where a common clock is distributed in the switch system since the phase of the optical packet is unknown and may vary from packet to packet. It is important to stress that the hybrid electro-optical packet switch has some advantages concerning the overhead in form of switching, clock recovery and skew compared to all-optical packet switches for WANs and LANs used in [3]. E.g. all-optical signals are generated locally and do not require any resynchronization to the time slots using fiber delayed loops (FDLs).

The contrast to the time slotted approach is to use a variable size optical packet. This would enable the switch to transport minimum size packets without wasting excess capacity in the optical packets. Unfortunately, it would require that the guard band used to reconfigure the switch should be reduced to a level much lower than the length of a minimum size IP packet in order to be efficient. Furthermore, it would require that the scheduler should produce results at a rate corresponding to transmission of minimum size packets. In this paper only fixed

0-7803-9569-7/06/$20.00 ©2006 IEEE

size time slot operation is considered.

Depending on the size of the optical slots (packets) on the optical backplane, and the size of the reconfiguration overhead, different scheduling and packet formatting strategies can be employed in order to get high throughput. Using a fixed length optical packet, gives two main possibilities when IP packets enter the individual VOQs:

1. The incoming IP packet is larger than the optical packet payload. The IP packet must be divided in multiple optical packets and send in different timeslots. Excess capacity can be used to other incoming IP packets.
2. The incoming IP packet is smaller than the optical packet payload. The IP packet is stored in the optical packet leaving excess capacity for future IP packets.

In order to support the above requirements, an optical packet format can be defined as follows: The incoming IP packet is divided into a number of fixed size segments. The segments are aggregated into the optical packet making it possible to transport multiple IP packets in one optical packet. Each segment contains a small header to keep information regarding the delineation of IP packets within the segment. This information is used on the egress line card to reassemble the IP packets. An optical packet containing three IP packets is illustrated in Fig. 2. The optimum number of segments in the optical packet depends on the optical packet size and the distribution of the incoming IP packet lengths. In this paper the segments are 64 bytes long, thus allowing minimum sized IP packets to be transported completely in one segment.

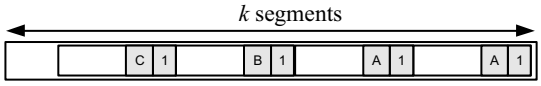

Figure 2 Optical packet format containing up to k segments. Here all segments are destined for line card output 1 but three different IP packets (A,B,C) have formed the segments. In this paper *k* equals 16.

II. SWITCH MODEL

The (simplified) line cards in Figure 1 segments the original IP packets into fixed length segments, and forwards them to the VOQs, where they are aggregated to form fixed length optical packets. When a segment enters an empty queue, a timer with value τ is started. Two different events can trigger the closing of an optical packet. Either it is closed because it is completely filled with segments, or due to the timer running out. The latter case gives rise to lower optical packet utilization, and therefore a switch fabric speedup is needed to compensate for this inefficiency in order to get 100% line speed throughput. The queue managers send information regarding the queue status to the central scheduler in each timeslot, and the central scheduler sends transmission acknowledges to the queue managers in return. Optical packets received at the output line card are stripped of any overhead, and the segments are used to reassemble the original IP datagram. Once an IP datagram is reassembled, it will be sent on the outgoing transmission line. The scheduler used in this evaluation is a modification of the iSLIP scheduler [4]. The modified scheduler i-ΔSLIP was proposed in [5] to enhance the throughput of iSLIP in switch fabrics with large round trip latency. Large round trip latency primary occurs due to the physical distance between line cards and switching fabric. In this paper an optical slot length of 1μs is selected. The optical packet used in the backplane contains 8192 bits, which is the equivalent of 16 segments each holding 64 bytes. If the external and internal interface line speeds are equal (e.g. 10Gb/s) a maximum total offered load would be 81.92% due to simple bandwidth limitation. Inefficiencies in scheduling and aggregation can actually lower this value significantly, which will be shown in the following section. Selecting a packet size of only 8192 bits at 10Gb/s line speed with 1μs timeslots, is of course very conservative, but it does not change the main results of this paper. The aggregation of segments in the VOQ of a line card is illustrated in Figure 3.

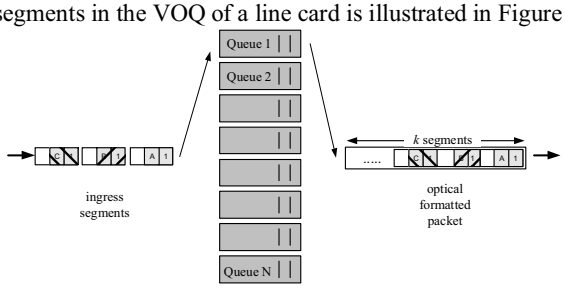

Figure 3 The VOQ system with aggregation of segments into optical packet.

Three segments arrive with the same output line card destination. The segments are stored in the optical formatted packet and eventually sent through the switch fabric. At the output line card, N output queues are needed, since all inputs might have un-assembled segments left. The fabric and line cards may run at different clock speeds, as illustrated by the dashed lines/box in Figure 1. This results in a speedup factor, which can be used to compensate for inefficiencies in the scheduling and aggregation processes. When a speedup factor larger than 1 is used, schedulers are required at the output line cards to resolve contention between different output queues.

Earlier studies [6] show that the selection of the timeout parameter τ has a high impact on the average delay that packets encounter in the aggregation stage. In the following section a number of simulations are carried out to illustrate the delay performance of the hybrid electro-optical packet switch with and without the adaptive timeout parameters.

III. SIMULATION RESULTS

The simulation model has been verified by a number of simulation studies used to compare with results presented in papers [4] and [5]. The results of this section are generated using a 32x32 hybrid electro-optical packet with no round trip delay. The scheduler is i-ΔSLIP with 5 iterations. TABLE I shows the simulation setup parameters for a number of scenarios with fixed timeout parameter τ.

TABLE I

SIMULATION SETUP	VALUE

Line cards	32
Optical packet length	1024 bytes
Segment length	64 bytes
Segments per optical packet	16
Time slot length	1us
Timeout parameter τ	1us-500us
Offered load (%)	1:10,20,30,40,50,60,70,75,80
Arrival process	On-Off model
On state distribution	Geometric
Off state distribution	Exponential
On state average length	16 segments (1 optical packet length)
Off state average length	Calculated from load

The result of simulating the various fixed timeout parameters can be seen in Figure 4.

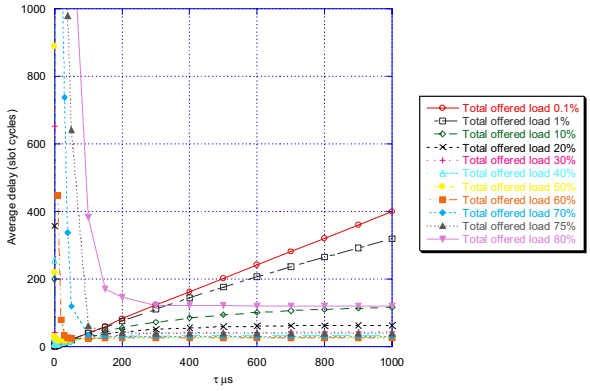

Figure 4 Simulation of the parameters listed in TABLE I.

It is observed that each of the different curves have a certain value of τ, that minimizes the average delay across the switch. The reason for this behavior is found in two different mechanisms. For values of τ lower than the optimum, optical packets are not filled completely, hence bandwidth in the switch fabric is wasted. For values of τ larger than the optimum, optical packets are closed due to full filling and thus the first segment in an optical packet has to wait for the last segment to arrive. It should be emphasized, that the average delay for all queues only is bounded by simulation time for values of τ under the optimum value. The reason for this is that the switch fabric is beyond its admissible capacity, resulting in growing queue lengths at the line cards. However there is one exception to the above statement: if the total offered load is below approximately 4.8%, average delay will not grow unbounded for low values of τ. The reason is that the traffic is admissible up to 77% total offered load (the graph with total offered load of 80% is only bounded by simulation time). The worst case scenario for low values of τ is when an incoming segment is encapsulated with 15 dummy segments forming an optical packet. Hence worst case admissible traffic is only approximately (77/16) %~4.8%.

A number of similar simulations have been carried out, only changing the average burst size, giving almost similar results. The optimum timeout values for these simulations can be plotted against the total offered load as shown in Figure 5. Doing this reveals that, if the total offered load can be measured accurately and fast, selecting close to optimum values of the timeout parameter is feasible. This is highlighted by the hysteresis function also drawn in Figure 5. A simpler two-step timeout parameter function could be used, but the hysteresis function is selected in order to avoid any oscillations between the two timeout values. Measuring the total offered load is done indirectly by measuring the mean inter arrival time of the segment arrival process in each VOQ. Measuring the mean inter arrival rate of the incoming segments can be highly fluctuating, so a low pass filtering is used to smooth the samples:

$$\Lambda_{avg}(n) = (1-\beta)\Lambda_{avg}(n-1) + \beta \cdot \Lambda(n)$$

Here β is the filter gain, $0 < \beta < 1$, determining the weight between old average $\Lambda_{avg}(n-1)$, and new sampled inter arrival rate $\Lambda(n)$. Selecting appropriate values of the filter gain is based on numerous simulation results, showing a good tradeoff between filter response and measurement accuracy. Here it is selected to 0.0025.

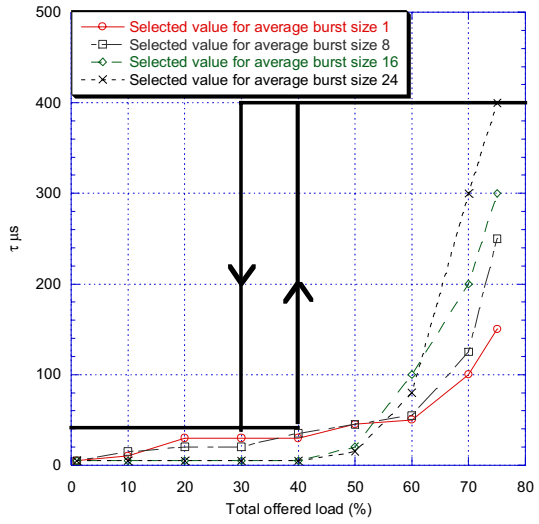

Figure 5 Optimum values of **τ** vs. total offered load

Running a simulation with the two step timeout function and the parameters from TABLE I (in addition to three other average burst sizes), gives the results in Figure 6. The results show that it is possible to achieve a low average delay for uniform and admissible traffic flows (i.e. below 77% total offered load). Here average delay is kept below 40us for all offered loads lower than 70%. Over 70% total offered load, delay is rising fast, but this is due to the well-known properties of the iSLIP scheduling mechanism. It does however not

indicate that other non-uniform traffic profiles will give bounded delay performance.

The techniques described above regulate τ on a per VOQ basis. This approach is acceptable when the traffic profile is uniform or close to uniform. When the traffic is not uniform, higher delay (and buffer overflow) can be experienced, if the input or output line card is oversubscribed. Oversubscribing the input is likely if one VOQ is heavily loaded while the rest is lighter loaded. The lighter loaded VOQs will produce more optical packet with low segment filling, resulting in oversubscription at the interface to the switch fabric.

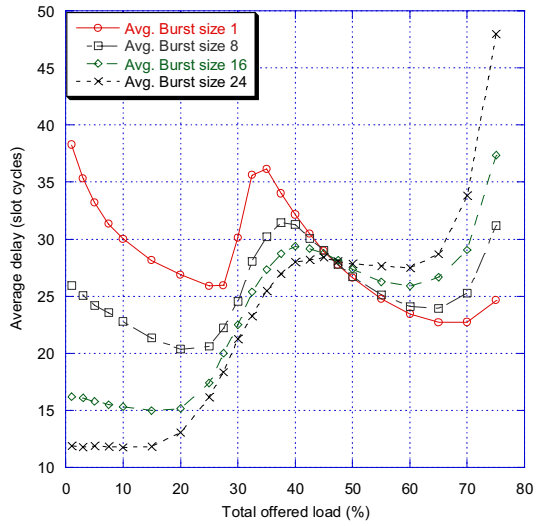

Figure 6 Average delay vs. total offered load for two stage timeout parameter with hysteresis

Providing bounded delay for all VOQs in the case of non-uniform traffic profiles, can be done by introducing two additional threshold levels. In the rest of this paper, we will refer to the threshold levels as flow control, due to its ability to reduce the generated number of empty segments in optical packets. The first flow control level is placed at the output of the input line cards. This will be termed input flow control in the following. The second flow control level is placed at the input of the output line cards. This level will be termed the output flow control in the following. The flow control is asserted in both cases if the optical packet slot load is higher than a certain threshold limit. For each input and output line card, measurements are made of the optical packet inter arrival rate. Measurements will be highly fluctuating (as with the segment measurements), so a low pass filter similar to the earlier presented one is used. The filtering gain is once again set to 0.0025 since it is a good trade off between filter response and measurement accuracy (found by modeling of the optical packet arrival process).

Based on the above description three levels of flow control (selecting τ) is used. For each VOQ only two values of τ exists (40us and 400us). Choosing which τ to use is based on the following three rules:

1. Each VOQ determine the mean inter arrival rate and choose τ based on the hysteresis depicted in Figure 5**.**
2. If an output line card is oversubscribed, it asserts flow control for all VOQs corresponding to that output line card. Hence the high value of τ is used once the flow control has propagated back to the input line cards.
3. If an input line card is oversubscribed, it asserts flow control for all VOQs in that line card. Hence the high value of τ is used almost instantaneous.

It is clear that traffic arrival patterns that tend to oversubscribe one or more output line cards, without oversubscribing the input line cards, is more buffer-consuming due to the fact that there is latency in the flow control signaling path. The VOQs must be able to accumulate incoming traffic, until the flow control is propagated back from the outputs. This is however not different from many other switching systems, utilizing multi level flow control. It is found through simulations that setting the input and output flow control threshold limits to 40% of the maximum load (e.g. 4Gb/s per interface) gives quite good delay performance for all investigated traffic arrival profiles. The simulation with parameters from TABLE I is repeated with the proposed flow control system. The result is illustrated in Figure 7.

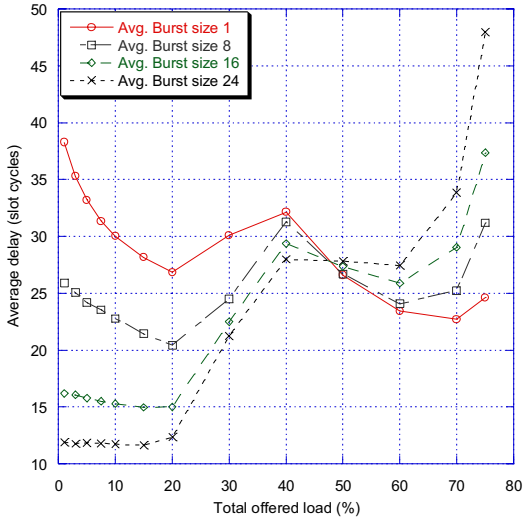

Figure 7 Average delay vs. total offered load with flow controls.

The results are very similar to the ones presented in Figure 6. This is obvious since the input and output flow control is asserted at the same thresholds as for rule number 1 in the case of uniform traffic arrival rates. Now a non-uniform distribution [7] is used to investigate the performance for non-uniform traffic arrivals:

$$\lambda_{i,j} = \begin{cases} \lambda\left(\omega + \dfrac{1-\omega}{N}\right) & \text{if } i = i \\ \lambda\dfrac{(1-\omega)}{N} & \text{otherwise} \end{cases}$$

$\lambda_{i,j}$ is the traffic intensity from input i to output j, N being the number of line cards. ω is the degree of non-uniformity. The following conditions apply:

$0 \le i, j < N$

$0 \le \omega \le 1$

The offered load per input and output port is admissible [5],[7] when: $0 \le \lambda \le 1$

$$\lambda_i = \sum_{j=o}^{N-1} \lambda i, j = \lambda_j = \sum_{i=0}^{N-1} \lambda_{i,j} = \lambda\left(\omega + N\frac{1-\omega}{N}\right) = \lambda$$

Simulating a number of different non-uniformity and average burst sizes gives the results shown in Figure 8-Figure 10.

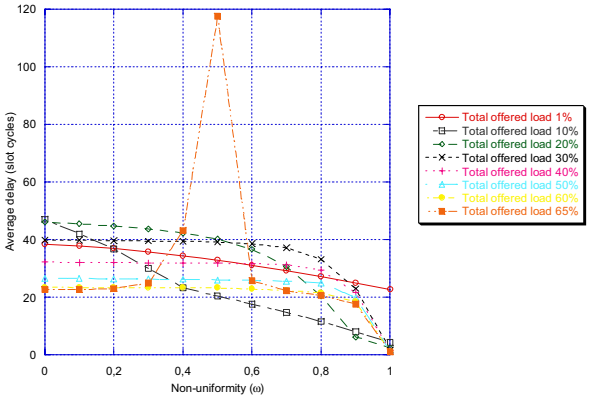

Figure 8 Average delay vs. non-uniformity and total offered loads for average burst size 1.

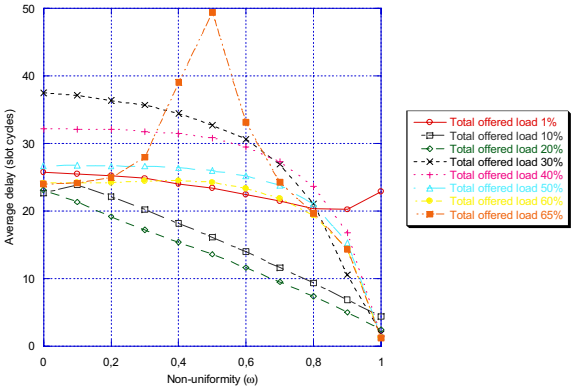

Figure 9 Average delay vs. non-uniformity and total offered loads for average burst size 8.

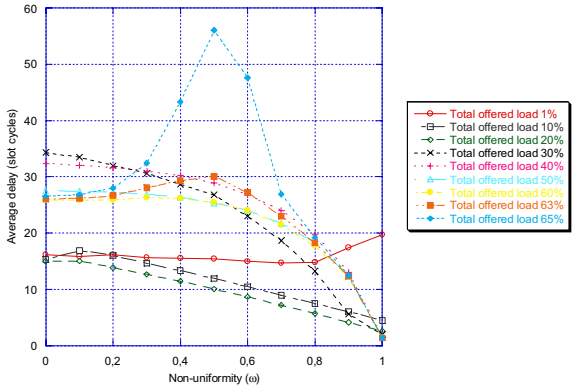

Figure 10 Average delay vs. non-uniformity and total offered loads for average burst size 16.

The simulations shows that bounded delay can be obtained, while providing low average delay, for a large number of non-uniform traffic arrival patterns and average burst sizes. The graphs in Figure 8-Figure 10 also show, that when the non-uniformity parameter is close to 0.5, and the total offered load is higher than 65%, a relatively large average delay is experienced. A number of simulations have shown, that a total offered load of 65% is very close to the maximum admissible offered load. Selecting a scheduler with higher throughput (and lower delay) for non-uniform traffic patterns, will result in better performance for the hybrid electro optical packet switch in this paper. With a speedup factor of around 1.55, 100% total offered load would be admissible for the non-uniform traffic rate arrivals used in this paper.

IV. Conclusion

Hybrid electro-optical packet switches are promising candidates for reaching aggregate switching bandwidth beyond 1Tb/s. This is primary due to reduction of component cost in form of conversion between optical and electrical domains. Furthermore there is a potential for reducing the power used in the switch backplane making packing of components denser.

In this paper a novel strategy for the packetisation process have been proposed. Measuring the segment inter arrival rates at each VOQ, and the optical slot inter arrival rates in the line cards, makes it possible to design a system which has high throughput and bounded delay. 100% throughput is possible using a speedup factor of approximately 1.55. The results in this paper can be further improved by choosing a scheduling algorithm that takes advantage of the long optical packet slot times.

References

[1] K. Kar, D. Stiliadis, T.V. Lakshman, L. Tassiulas, "Scheduling Algorithms for Optical Packet Fabrics" IEEE Journal on selected areas in communication, vol. 21, NO. 7, September 2003

[2] X. Li, M. Hamdi, On Scheduling Optical Packet Switches with Reconfiguration Delay" IEEE Journal on selected areas in communication, vol. 21, NO. 7, September 2003

[3] L. Dittmann, C. Develder, F. Neri, F. Callegati, Menber, IEEE, W. Koerber, A. Stavdas, M. Renaud, A. Rafel, J. Solé-Pareta, W. Cerroni,

N. Legilou, L. Dembeck, B. Mortensen, M. Pickavet, N. Le Sauze, M. Mahony, B. Berde, and G. Eilenberger, "The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching" IEEE Journal on selected areas in communication, vol 21, NO. 7, Sebtember 2003

[4] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches". IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 7, NO. 2, APRIL 1999, 188-200.

[5] C. Minkenberg, "Performance of i-SLIP Scheduling with Large Round-Trip Latency". Workshop on High Performance Switching and Routing, Torino, Italy, June 24-27, 2003, 49-54.

[6] M. S. Berger, B. B. Mortensen, V. B. Iversen, R. Jociles-Ferrer, Evaluation of Delay Bound for QoS provisioning in Optical Packet Network Interface, 7th WSEAS International Conference on Communications, invited paper, Corfu, Greece, 2003, 103-108.

[7] R. Rojas-Cessa, E.Oki,Z.Jing, H.J.Chao, CIXB-1 Combined input-one-cell-crosspoint buffered, Workshop on High performance Switching and Routing, Dallas, Tx, USA, May 2001, 324-239.